Do Citations Improve Grounding? A Controlled Study of Local RAG for Scientific Question Answering

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) is often presented as a path toward trustworthy question answering, especially when answers include citations. We test this assumption in local scientific question answering on QASPER, where answers should be supported by evidence from research papers. We build a fully local RAG pipeline and run controlled ablations over chunking, sparse and dense retrieval, hybrid reciprocal-rank fusion, reranking, and citation-forcing prompts. Our strongest baseline uses section-aware chunks, hybrid retrieval, reranking, and ordinary answer generation, reaching token F1 0.3238 and supported-answer rate 0.7652 on the QASPER validation set; a Llama-3-8B generator sanity check is effectively tied, while Llama citation forcing repeats the grounding drop. Citation forcing reaches a high citation rate (0.9274), but lowers supported-answer rate to 0.5960 and yields low NLI citation precision (0.1416). Bootstrap intervals show that citation forcing is not clearly worse on token F1 but is clearly worse on support, and untuned PeerQA and QASA transfer checks reinforce that prompt-only citation is unreliable. Manual audits sharpen the contrast: only 38.0% of prompted citations are directly supported, while selective ClaimAlign v2 citations are directly supported 88.0% of the time but appear on only 26.97% of answers. These results suggest that, for local scientific QA, citations can make outputs look grounded without making them faithfully supported unless citation production is tied to claim-level evidence alignment.

1 Introduction

Scientific question answering is a demanding test bed for grounded language generation: a useful answer should be not only fluent, but also traceable to specific evidence in a paper. Retrieval-augmented generation (RAG) gives language models access to document evidence (Lewis et al., 2020), and citation-aware prompting asks models to expose that evidence to users. However, visible citations are not the same as faithful attribution. A model can cite frequently while citing the wrong passage, or cite a relevant passage while making a claim that is not actually supported by it.

We study this issue in a local setting. Our pipeline uses no external APIs and is designed to run on commodity shared compute, which makes the setting relevant for privacy-sensitive scientific and institutional documents. We use QASPER (Dasigi et al., 2021), a scientific QA benchmark with paper questions, gold answers, and gold evidence spans. The gold evidence lets us evaluate retrieval, answer quality, and citation grounding separately rather than collapsing all system behavior into one text-similarity score. We also run no-retuning transfer checks on PeerQA (Baumgärtner et al., 2025) and QASA (Lee et al., 2023) to test whether the citation-forcing pattern is specific to QASPER.

The paper asks two questions:

1. How much does answer quality depend on retrieval and chunking choices, compared with generator prompting?

2. Does citation forcing improve groundedness, or does it mainly make answers appear better attributed?

Our answer is diagnostic and negative: stronger retrieval and section-aware chunking improve answer support, while citation forcing increases visible citation behavior but reduces automatic and human grounding scores. We then test a selective post-hoc alternative, ClaimAlign, which treats citation as a separate evidence-alignment decision rather than as a formatting instruction to the generator. This makes citation forcing a risky trust signal when it is not paired with stronger evidence alignment.

Our contributions are:

- a controlled local RAG study on QASPER separating chunking, retrieval, reranking, and prompting effects;
- evidence that citation forcing can reduce grounding despite high citation rate;
- a metric analysis combining token F1, evidence-support rate, citation precision, NLI citation precision, bootstrap intervals, and human judgments;
- ClaimAlign, a selective post-hoc citation-alignment method that tests claim-level retrieval and abstention;
- no-retuning PeerQA and QASA transfer stress tests outside the main dataset;
- a reproducible experimental pipeline for local scientific QA.

2 Related Work

RAG was introduced to improve knowledge-intensive NLP by conditioning generation on retrieved evidence (Lewis et al., 2020). Dense passage retrieval made neural retrieval a practical alternative to sparse lexical matching (Karpukhin et al., 2020), while reciprocal-rank fusion provides a simple way to combine ranked retrieval lists (Cormack et al., 2009). Scientific QA benchmarks such as QASPER evaluate whether systems can answer questions grounded in research papers (Dasigi et al., 2021); PeerQA adds scientific QA questions derived from peer-review needs (Baumgärtner et al., 2025), and QASA targets deeper reasoning over scientific articles (Lee et al., 2023). Recent systems extend retrieval to broader scientific literature search and synthesis (Ajith et al., 2024; Lálá et al., 2023).

Citation-aware generation is related to attribution and faithfulness evaluation. ALCE evaluates whether language models can answer with citations (Gao et al., 2023), and AIS frames attribution as whether generated claims are supported by identified sources (Rashkin et al., 2023). Recent work further distinguishes citation correctness from citation faithfulness: a cited passage can support a claim while not reflecting the evidence the model actually used (Wallat et al., 2024). Other work compares generation-time and post-hoc citation, finding that retrieval quality is a central driver of attribution quality (Saxena et al., 2025). Stronger citation systems often add training or structure, such as fine-grained grounded quotes (Huang et al., 2024), long-context citation supervision (Zhang et al., 2024), or answer decomposition for post-hoc attribution (Ramu et al., 2024). Our work follows this evaluation direction, but focuses on a controlled local QA setting where gold evidence spans make retrieval and citation failures directly measurable. Unlike training-heavy citation systems, our method isolates what can be gained by decoupling local answer generation from selective citation alignment.

3 Task and System

Given a paper p and question q, the system retrieves a ranked list of chunks $c_{1:k}$ from p and generates an answer $\hat{y}$, optionally with citations back to retrieved chunks. A prediction is useful only when both the answer and its evidence are correct enough for the answer to be grounded.

Chunking. We normalize QASPER papers into an internal schema and compare fixed-size chunks, sliding-window chunks, and section-aware chunks. Section-aware chunks preserve paper structure, which is useful because many scientific questions refer to methods, datasets, or results that are localized by section.

Retrieval. We compare BM25 (Robertson and Zaragoza, 2009), dense retrieval with BGE-small embeddings (Xiao et al., 2023), hybrid retrieval with reciprocal-rank fusion, and hybrid retrieval followed by a cross-encoder reranker (Nogueira and Cho, 2019). Hybrid retrieval scores chunk c as

$$\mathrm{RRF}(c) = \sum_{m \in \{\mathrm{bm25}, \mathrm{dense}\}} \frac{1}{k + \mathrm{rank}_m(c)}.$$

Generation. The primary controlled full-validation generation experiments use Phi-3.5-mini-instruct (Microsoft, 2024). We compare a baseline prompt against citation forcing, where the model must attach bracketed chunk citations to answer claims. To test whether the pattern depends on Phi, we also run Llama-3-8B follow-ups for the strongest baseline and citation-forcing settings.

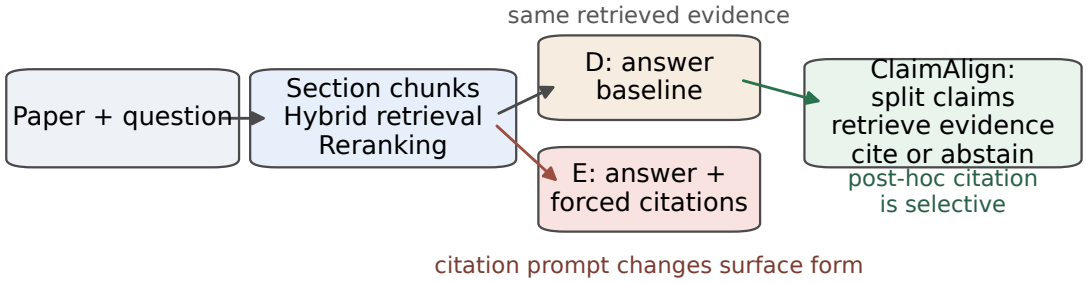

Figure 1: Experimental setup. D and E use the same retrieved evidence, but E asks the generator to produce citations inline. ClaimAlign keeps D's answer fixed and makes citation a separate claim-level decision with abstention.

Claim-level post-hoc citation. Citation forcing asks the generator to answer and cite in one step, which makes it hard to tell whether a citation reflects actual evidence alignment or only post-hoc formatting. We therefore add ClaimAlign, a post-hoc citation method that keeps the answer text fixed and decides citations separately. ClaimAlign splits $\hat{y}$ into sentence-level claims h_i, builds a claim-conditioned query from q and h_i, scores candidate chunks, and cites the best chunk only if its score exceeds a threshold τ:

$$e_i^* = \arg\max_e s(q, h_i, e),$$
$$\mathrm{cite}(h_i) = \mathbf{1}[\max_e s(q, h_i, e) \geq \tau].$$

V1 scores only the chunks retrieved for the original answer. V2 retrieves up to 20 claim-conditioned candidates from all chunks in the same paper before scoring. We test both a dependency-free lexical score,

$$s_{\mathrm{lex}} = p(h_i, e)\left(0.8R(T_{h_i}, T_e) + 0.2R(T_q, T_e)\right),$$

where R is token-overlap recall and p down-weights one-token matches, and an NLI score from a cross-encoder entailment model. The important behavior is abstention: a claim receives no citation when no candidate clears the threshold. This converts citation from an all-or-nothing generation style into a selective prediction problem with an explicit coverage–precision tradeoff.

4 Evaluation

Retrieval is evaluated with Recall@5, MRR@10, and evidence hit rate. Answer quality is evaluated with token F1 and BERTScore, but we treat BERTScore only as a secondary paraphrase signal because it can overestimate semantic similarity for unsupported answers.

For grounding, we report supported-answer rate and a hallucination proxy. Supported-answer rate is the mean of `answer_supported`: a prediction is supported if its cited chunks, or its retrieved chunks when no citations are produced, contain a gold-evidence chunk. Correct `unanswerable` predictions also count as supported; when QASPER provides no gold evidence text, we fall back to nonzero token F1 as a weak support signal. The hallucination proxy is

$$\mathbf{1}[\mathrm{tokenF1} = 0 \vee \texttt{answer_supported} = 0],$$

so it is not simply $1 - \mathrm{Supported}$.

For citation outputs, citation precision is the fraction of cited chunks that match gold evidence. For selective citation systems, we also report conditional citation precision over produced citations, excluding abstained questions. NLI citation precision is the fraction of cited claim sentences that are entailed by at least one cited chunk. In human evaluation, reviewers score answer correctness and grounding on a 0–2 scale, where 2 is clearly correct or directly supported, 1 is partially correct or partially supported, and 0 is wrong or unsupported. We also manually audit 50 cited claim–chunk pairs from the citation-forcing run and 50 produced ClaimAlign v2 citation pairs with two independent reviewers, using the same 2/1/0 direct/partial/unsupported support labels and then adjudicating disagreements.

5 Experiments

The main ablation consists of five controlled Phi settings:

Method	Recall@5	MRR@10	Hit
BM25	0.5769	0.4110	0.7152
Dense	0.5886	0.4820	0.7238
Hybrid	0.6313	**0.5097**	**0.7778**
Hybrid + rerank	**0.6410**	0.5027	0.7702

Table 1: Validation retrieval results on section-aware chunks.

- A: fixed chunks, dense retrieval, baseline prompting;
- B: fixed chunks, hybrid retrieval, baseline prompting;
- C: section-aware chunks, hybrid retrieval, baseline prompting;
- D: section-aware chunks, hybrid retrieval, reranking, baseline prompting;
- E: the same retrieval pipeline as D, but with citation forcing.

We also compare two full-validation baselines: parametric-only generation with no retrieval, and BM25-only retrieval. To check whether the main generator effects are Phi-specific, we rerun D and E with Llama-3-8B on full validation and screen Mistral-7B under D on 100 questions. A targeted human evaluation compares D and E on 50 questions with two reviewers. The reviewer rubric scores correctness, grounding, citation quality, and pairwise preference. Finally, ClaimAlign starts from D's answer text and tests whether selective post-hoc citation can repair citation forcing without changing the answer generator. We report threshold sweeps in the supplement and include a strict V2 setting that removes D's original support context to separate gains from new claim-conditioned retrieval versus gains from retaining the original retrieved evidence. As transfer checks, we process PeerQA into the same schema and run all 210 test questions; for QASA, we use ar5iv body text when available, append released evidence contexts as a backstop, and run a fixed 210-question test sample spanning 96 papers. Neither transfer setting is retuned.

6 Results

Table 1 shows that hybrid retrieval gives the best overall retrieval balance. Reranking improves Recall@5 slightly but does not dominate all retrieval metrics, suggesting that top-of-list reranking is useful but not enough by itself.

The main result is that retrieval and chunking matter more than citation forcing. Moving from fixed chunks to section-aware chunks improves token F1 and support. Adding reranking gives the strongest overall baseline (D). Parametric-only generation collapses, confirming that retrieval is essential. Citation forcing (E) nearly matches D on token F1 and BERTScore, but it raises the hallucination proxy and sharply lowers supported-answer rate. Paired bootstrap confidence intervals with 10,000 resamples show that the token F1 delta is small and uncertain (-0.0023, 95% CI $[-0.0148, 0.0099]$), while the support drop is large (-0.1692, 95% CI $[-0.1930, -0.1463]$) and the hallucination increase is clear ($+0.0985$, 95% CI $[0.0736, 0.1224]$). Its high citation rate therefore does not imply faithful attribution.

The generator checks reduce two single-model concerns. First, the non-citation baseline is not specific to Phi: under D, Llama-3-8B reaches token F1 0.3249, BERTScore 0.8801, hallucination 0.4537, and support 0.7552 on the full validation set, effectively tying Phi's F1 while slightly trailing its support. Second, citation forcing also fails under Llama. Full-validation Llama E reaches token F1 0.2598, hallucination 0.6706, support 0.5602, citation rate 0.8896, citation precision 0.3987, and NLI citation precision 0.1493. Relative to Llama D, support drops by 0.1950 and hallucination rises by 0.2169. A 100-question screening run with Mistral-7B under D underperforms on token F1 (0.2878), although its support remains high (0.82). These generator swaps do not prove full model generality, but they make the citation-forcing failure less likely to be a Phi-only prompting artifact.

The no-retuning transfer runs give harder stress tests. On PeerQA, D reaches token F1 0.1500 and support 0.1238; citation forcing raises citation rate to 0.9571 but drops token F1 and support to 0.0823 and 0.0333. On the QASA sample, D reaches token F1 0.1855, BERTScore 0.8522, and support 0.8476; citation forcing drops to token F1 0.0138 and BERTScore 0.8008, raises hallucination from 0.3857 to 0.9667, and cites only 7.62% of answers. QASA shows a different failure mode: the prompt can fail to elicit citations while still damaging answers. We treat both as transfer stress tests, not tuned benchmark submissions; full details are in the supplement.

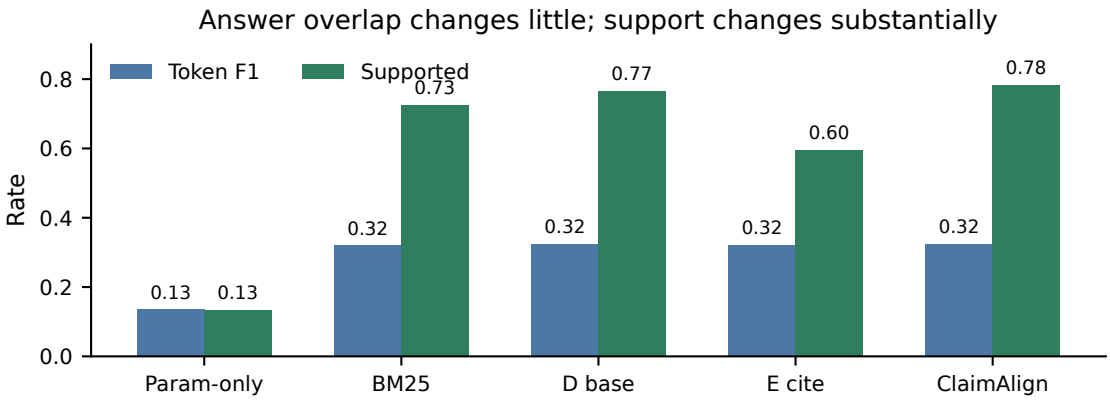

Figure 2: Answer-overlap metrics obscure grounding differences. Citation forcing stays close to D on token F1, but support drops sharply; ClaimAlign keeps D's answer text and selectively restores support by abstaining on weak citations.

Setting	Questions	Token F1	BERTScore	Halluc.	Supported	Cite/NLI Prec.
A: fixed + dense + base	1005	0.3059	0.8791	0.5134	0.6886	–
B: fixed + hybrid + base	1005	0.3063	0.8782	0.5104	0.6716	–
C: section + hybrid + base	1005	0.3197	0.8815	0.4527	**0.7741**	–
D: section + hybrid_rerank + base	1005	**0.3238**	0.8817	**0.4478**	0.7652	–
E: section + hybrid_rerank + citation	1005	0.3215	**0.8819**	0.5463	0.5960	0.3230 / 0.1416
P: parametric-only	1005	0.1350	0.8329	0.8657	0.1343	–
M: BM25-only	1005	0.3207	0.8813	0.4716	0.7254	–

Table 2: Full-validation answer results. Citation rate for E is 0.9274.

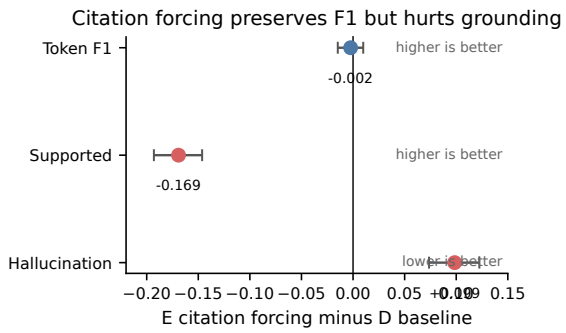

Figure 3: Paired bootstrap deltas for E minus D. The lexical answer-quality difference overlaps zero, while grounding metrics move decisively in the wrong direction.

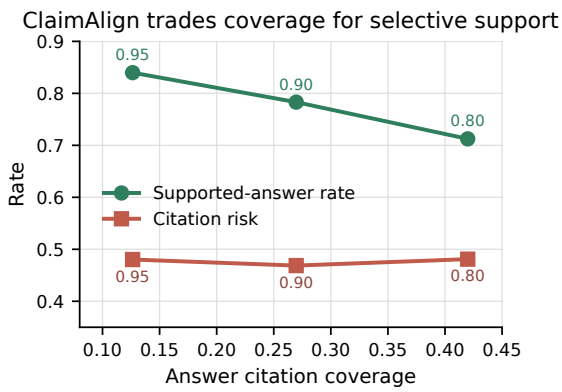

Figure 4: ClaimAlign v2 risk/coverage tradeoff. Stricter citation thresholds reduce answer citation coverage; support improves mainly because abstained answers retain broader retrieved context, while automatic citation risk remains roughly flat.

ClaimAlign helps, but remains limited. It reuses D's answer text, so token F1 remains 0.3238. V1 NLI-gated ClaimAlign raises supported-answer rate from 0.5960 to 0.6866 and lowers the hallucination proxy from 0.5463 to 0.5114, but cites only 19.9% of answers and has low gold-evidence citation precision (0.0766). V2 is stronger: lexical claim-conditioned retrieval with a 0.90 threshold reaches support 0.7831, hallucination 0.4348, citation rate 0.2697, and conditional citation precision 0.5314 over produced citations. V2 with NLI scoring is more conservative but weaker on this metric stack, reaching support 0.7383, hallucination 0.4816, citation rate 0.2557, and conditional citation precision 0.3268.

Figure 4 shows why selective citation must be interpreted as a coverage–risk tradeoff. Raising the lexical threshold from 0.80 to 0.95 cuts an-

Metric	Base	Citation
Correctness (0–2)	1.110	0.970
Grounding (0–2)	1.250	1.080
Citation quality (0–2)	–	0.880
Majority preference	12	8
Tie questions	30	
Preference agreement	56.0%	

Table 3: Targeted human evaluation on 50 questions with two reviewers.

Audit metric	Prompt cite	ClaimAlign v2
Exact agreement	76.0%	86.0%
Weighted κ	0.7588	0.7076
Direct support	38.0%	88.0%
Partial-or-better support	76.0%	100.0%

Table 4: Manual citation audits on 50 cited claim–chunk pairs. ClaimAlign v2 is the selective lexical 0.90 setting; its support rates are conditional on produced citations.

swer citation coverage from 0.4199 to 0.1264 and claim coverage from 0.5262 to 0.1584, while automatic citation risk stays near 0.47–0.48. A strict V2 run that drops D's original context falls to support 0.6846, so the support gain partly depends on retaining the original retrieved evidence. Selective post-hoc attribution is therefore better than forcing citations everywhere, but citation quality remains sensitive to the scorer and abstention threshold.

Human evaluation supports the automatic grounding story but adds nuance. Reviewers give the baseline higher mean correctness and grounding scores. On the 100 reviewed system-question rows, token F1 aligns more closely with human correctness than BERTScore does (Pearson $r = 0.865$ vs. 0.339), consistent with the concern that BERTScore can overstate answer quality. However, majority preference is often tied: 30 of 50 questions have no majority winner, while 12 favor the baseline and 8 favor citation forcing. Citation answers are therefore not uniformly rejected; the problem is inconsistency, where citations sometimes make a good answer easier to trust but often provide weak or unsupported attribution.

Table 4 gives a claim-level check on this pattern. For prompted citations, only 38.0% of cited claim–chunk pairs are directly supported after adjudication, even though 76.0% receive at least partial support. A separate audit of produced ClaimAlign v2 citations is much stronger: 88.0% are directly supported and all 50 sampled citations receive at least partial support. This result is conditional on selective citation, so it should not be read as coverage; instead, it shows that abstaining and retrieving evidence per claim can make produced citations substantially more trustworthy than prompt-forced citations.

7 Discussion

These results argue against treating citation prompts as a cheap substitute for evidence alignment. Citation forcing changes the surface form of the answer, but it does not fix retrieval misses and introduces unsupported or only weakly supported citations. Claim-conditioned post-hoc alignment helps because it can abstain and search for evidence at the claim level, yet its gains depend on the support scorer and on preserving the broader retrieved context. For local scientific QA, the more reliable intervention is still retrieval quality: better chunking, hybrid retrieval, and reranking reduce the chance that the generator must answer from incomplete or misleading evidence.

Together, the bootstrap intervals, human metric alignment, and citation audit make the negative result more specific: citation-looking answers can be less grounded than ordinary answers under the same retrieved evidence, BERTScore can overstate wrong yes/no answers, and automatic NLI checks catch only part of the citation-support problem.

8 Conclusion

In local scientific QA, retrieval improvements are more dependable than citation-forcing prompts. Section-aware chunks, hybrid retrieval, and reranking improve support, while citation forcing increases visible citation behavior but reduces groundedness. Claim-conditioned post-hoc alignment is a more promising direction than prompt-only citation forcing, but the next step is not to make citations more frequent; it is to align generated claims to evidence more precisely.

9 Limitations

This study uses QASPER as the primary dataset for the controlled A–E ablations, with PeerQA and QASA included only as untuned transfer stress tests. The main generator is Phi, with Llama follow-ups for D and E rather than a full multi-model grid. Additional tuned scientific QA evaluations would still be needed to establish general-

ity across domains and question styles. The human evaluation and citation audits are small, so the manual findings should be read as diagnostic rather than definitive. ClaimAlign v2 also relies on sentence-level claims and single-chunk support decisions, which can underserve multi-hop claims or claims that require synthesizing several passages. The local compute setting is intentional, but it means the results should be read as evidence about practical local RAG systems rather than as a claim about frontier proprietary models.

10 Ethical Considerations

This work studies scientific question answering over public research-paper benchmarks. The main risk is overtrust: visible citations can make an answer appear more reliable than it is. We therefore report support, hallucination, citation precision, and manual audit results separately, and we recommend selective citation with abstention rather than forced citation in high-stakes settings. The human evaluation used project reviewers to judge model outputs and cited evidence; no new private user data was collected.

References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. LitSearch: A retrieval benchmark for scientific literature search. In *Proceedings of EMNLP*. ArXiv:2407.18940.

Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. 2025. PeerQA: A scientific question answering dataset from peer reviews. In *Proceedings of NAACL*. ArXiv:2502.13668.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of SIGIR*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of NAACL*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of EMNLP*. ArXiv:2305.14627.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024. Learning fine-grained grounded citations for attributed large language models. In *Findings of ACL*. ArXiv:2408.04568.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*.

Jan Lálá, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. PaperQA: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-In Lee, and Moontae Lee. 2023. QASA: Advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19036–19052.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2005.11401.

Microsoft. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage reranking with BERT. *arXiv preprint arXiv:1901.04085*.

Pritika Ramu, Koustava Goswami, Apoorv Saxena, and Balaji Vasan Srinivasan. 2024. Enhancing post-hoc attributions in long document comprehension via coarse grained answer decomposition. In *Proceedings of EMNLP*. ArXiv:2409.17073.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4).

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Yash Saxena, Raviteja Bommireddy, Ankur Padia, and Manas Gaur. 2025. Generation-time vs. post-hoc citation: A holistic evaluation of LLM attribution. *arXiv preprint arXiv:2509.21557*.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2024. Correctness is not faithfulness in RAG attributions. *arXiv preprint arXiv:2412.18004*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packed resources for general Chinese embeddings. *arXiv preprint arXiv:2309.07597*.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. *arXiv preprint arXiv:2409.02897*.